Econometria

Multicolinearidade

Multicolinearidade

- Muitas variáveis econômicas podem caminhar juntas de maneira sistemática.
- Tal fenômeno é denominado colinearidade e, quando consideramos diversas variáveis, multicolinearidade.
- Originalmente, significava a existência de uma relação linear "perfeita". entre algumas ou todas as variáveis explicativas do modelo de regressão.
- Atualmente, o termo Multicolinearidade é usado no sentido mais amplo em que as variáveis explicativas estão intercorrelacionadas, mas não perfeitamente.

MULTICOLINEARIDADE

Uma das hipóteses do Modelo Clássico de Regressão Linear (MCRL) é a de que não existe multicolinearidade perfeita entre as variáveis explicativas.

Conceitualmente, a multicolinearidade ocorre quando há uma relação linear perfeita entre os regressões, de tal forma que

$$\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_k x_k = 0$$

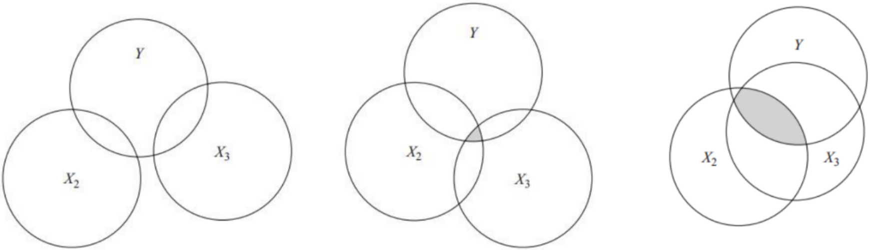
Menos estritamente, o termo "*multicolinearidade*" também tem sido empregado mesmo quando as relações entre os regressões não são perfeitas:

$$\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_k x_k + u_i = 0$$

Exemplo Multicolinearidade Perfeita x Multicolinearidade Imperfeita

Multicolinearidade

Diagrama de Ballentine:



(*a*) Ausência de colinearidade

(*b*) Baixa colinearidade

(*e*) Colinearidade muito alta

ALGUMAS FONTES DE MULTICOLINEARIDADE

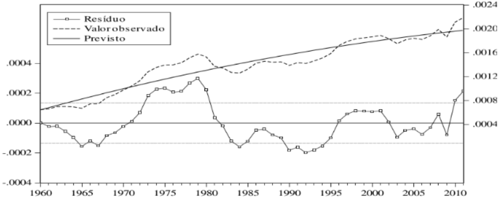
a) **Coleta de dados**:

A amostra pode ser escolhida de tal forma que os regressores apresentem grande relação.

Ex: Relação entre renda e riqueza (função de consumo)

a) **Especificação do modelo**:

A inclusão de termos polinomiais, principalmente quando a amplitude da variável x é pequena.



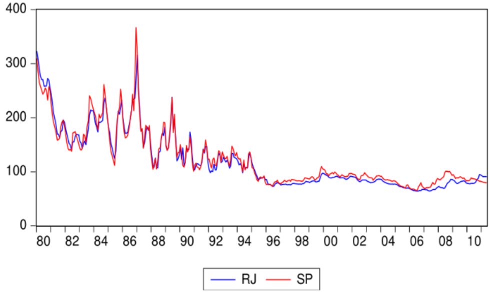
ALGUMAS FONTES DE MULTICOLINEARIDADE

c) **Modelo sobredeterminado**

Número de parâmetros maior que o número de observações.

d) **Existência de tendência comum**

Todas as variáveis aumentam ou diminuem ao longo do tempo.



Estimação na Presença de Multicolinearidade Perfeita

CONSIDERE UM MODELO DE REGRESSÃO COM DUAS VARIÁVEIS EXPLICATIVAS.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

USANDO MQO OBTEMOS

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$

LEMBRANDO:

$x_i = X_i - \bar{X}$

$y_i = Y_i - \bar{Y}$

$\sum x_i^2 = \sum (X_i - \bar{X})^2$

$\sum x_i y_i = \sum x_i (Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y} \sum x_i$

Estimação na Presença de Multicolinearidade Perfeita

A LETRA MINÚSCULA INDICA OS DESVIOS EM RELAÇÃO AOS VALORES MÉDIOS.

SUPONHA QUE $X_{3i} = \lambda X_{2i}$, EM QUE λ É UMA CONSTANTE DIFERENTE DE ZERO. ASSIM,

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} = \frac{0}{0}$$

QUE É UMA EXPRESSÃO INDETERMINADA. LEMBRANDO QUE $\hat{\beta}_2$ NOS DÁ A VARIAÇÃO DO VALOR MÉDIO DE Y QUANDO X_2 VARIA POR UMA UNIDADE, MANTENDO X_3 CONSTANTE. CONTUDO, SE X_2 E X_3 FOREM PERFEITAMENTE COLINEARES, NÃO HAVERÁ COMO MANTER X_3 CONSTANTE.

Estimação na Presença de Multicolinearidade Perfeita

QUANDO X_2 VARIA, X_3 TAMBÉM VARIA PELO FATOR λ. ISTO É, NÃO HÁ COMO DISTINGUIR AS INFLUÊNCIAS DE X_2 E X_3 DE FORMA SEPARADA NA AMOSTRA DADA.

VAMOS AGORA SUBSTITUIR $X_{3i} = \lambda X_{2i}$, NA EQ. DE REG.

Estimação na Presença de Multicolinearidade Perfeita

Estimação na Presença de Multicolinearidade Perfeita

Intuitivamente, se duas variáveis explicativas possuem relação perfeita, não há como separar sua influência individual sobre y, o que é um problema grave.

Ex: $$\begin{cases} y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \\ x_{3i} = \lambda x_{2i} \end{cases}$$

MULTICOLINEARIDADE ALTA, MAS IMPERFEITA

A multicolinearidade perfeita é um caso extremo. Em geral, temos:

$$\lambda_1 x_1 + \lambda_2 x_2 \boxed{+ u_i} = 0$$

A introdução do componente aleatório torna possível a estimação dos parâmetros.

É realmente necessária a suposição de Multicolinearidade?

- Em caso de multicolinearidade perfeita, os coeficientes de regressão serão indeterminados e seus erros padrão, infinitos.
- Quando a multicolinearidade é imperfeita, embora determinados, os coeficientes de regressão poderão possuir erros padrão altos, de forma que os coeficientes não poderão ser estimados com grande precisão.
- A multicolinearidade não viola nenhuma das suposições do modelo clássico de regressão.
- Mesmo na presença de multicolinearidade teremos estimadores BLUE.

É realmente necessária a suposição de Multicolinearidade?

- O fato dos estimadores serem BLUE não significa que a variância é pequena em relação ao valor do estimador e sim que entre todos estimadores não viesados, o estimador de MQO tem variância mínima.

- O único problema é que devido aos erros padrões grandes das estimativas, teremos estimação menos precisa.

- Contudo, outras características da amostra também podem gerar essa imprecisão:

  I. Amostras pequenas

  II. Variáveis explicativas com pequena variação.

CONSEQUÊNCIAS

Em resumo, a multicolinearidade é um problema tão grave quanto baixa variabilidade de x e amostras pequenas (micronumerosidade). Trata-se de um problema amostral.

Mesmo com multicolinearidade imperfeita, os parâmetros continuam não tendenciosos. Na verdade, viés é uma propriedade de amostras repetidas, não de uma amostra particular.

A variância continua sendo mínima (estimador é eficiente).

O aumento do tamanho da amostra pode reduzir o problema de multicolinearidade.

CONSEQUÊNCIAS

O estimador é BLUE, mas com alta variância

Razões t "insignificantes"

$$t = \frac{\hat{\beta}_2}{ep\left(\hat{\beta}_2\right)}$$

Como a variância dos estimadores será mais alta na presença de multicolinearidade, as estatísticas t tendem a ser baixas.

CONSEQUÊNCIAS

Intervalos de confiança mais amplos

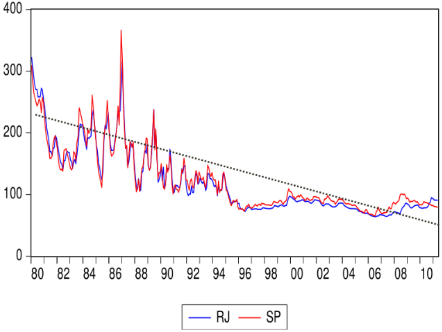
$$pr\left[\hat{\beta}_2 - \boxed{ep\left(\hat{\beta}_2\right)} \cdot t_{\alpha/2} \le \beta_2 \le \hat{\beta}_2 + \boxed{ep\left(\hat{\beta}_2\right)} \cdot t_{\alpha/2}\right] = 1 - \alpha$$

A variância do estimador será maior e, consequentemente, seu erro padrão. Por isso, os intervalos na presença de multicolinearidade tendem a ser grandes.

CONSEQUÊNCIAS

Alto valor de R^2 com "t's" insignificantes

Os testes individuais podem indicar insignificância dos parâmetros, mas o R^2 pode ser alto, sobretudo em modelos de séries temporais – as variáveis podem compartilhar a mesma tendência.



CONSEQUÊNCIAS

Sensibilidade dos estimadores de MQO e dos erros padrão a pequenas alterações nos dados

São problemas de micronumerosidade (amostras pequenas). Os erros padrão podem ser sensíveis e elevados em razão do baixo tamanho amostral.

RESUMO: Consequências da Multicolinearidade

- Estimação imprecisa dos parâmetros devido às grandes variâncias dos estimadores de MQO.
- Intervalos de confiança dilatados (mais amplos).
- Os testes *t* podem levar a conclusões de que as estimativas são insignificantes(não são significativamente diferentes de zero).
- Valores elevados do coeficiente de determinação, indicando poder explicativo significativo do modelo como um todo.
- Os estimadores e seus erros padrão podem ser muito sensíveis a pequenas alterações nos dados (acréscimo ou supressão de observações ou variável aparentemente insignificante.).

EXEMPLO

Considere as variáveis: Consumo (Y), Renda (X2) e Riqueza (X3). Considere a equação de regressão para explicar o consumo. É possível detectar algum problema de Multicolinearidade?

$$\begin{array}{llll}
\hat{Y}_i = & 24{,}7747 & + \; 0{,}9415X_{2i} & - \; 0{,}0424X_{3i} \\
 & (6{,}7525) & (0{,}8229) & (0{,}0807) \\
t = & (3{,}6690) & (1{,}1442) & (-0{,}5261) \\
\end{array}$$

$$R^2 = 0{,}9635 \qquad \bar{R}^2 = 0{,}9531 \qquad \text{gl} = 7$$

Os Componentes da Variância de MQO e a Multicolinearidade

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} x_j^2 (1 - R_j^2)}$$

VARIÂNCIA DO ERRO: σ^2

- Um σ^2 maior significa variâncias maiores dos estimadores de MQO: Mais ruído na equação torna mais difícil estimar o efeito de qualquer variável independente.
- Como σ^2 refere-se a população não tem nenhuma relação com o tamanho da amostra.
- Uma maneira de reduzir a variância do erro é adicionar mais variáveis explicativas. Mas nem sempre é possível encontrar variáveis adicionais que afetem Y.

Os Componentes da Variância de MQO e a Multicolinearidade

$SQT_j = \sum_{i=1}^{n} x_j^2 = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$

- Quanto maior a variação total em X_j, menor é a $Var(\hat{\beta}_j)$.
- Quanto mais dispersa for a amostra das variáveis independentes, mais fácil será descrever a relação entre $E(Y|X)$ e X, isto é estimar β.
- Se há pouca variação nos X pode ser difícil estabelecer com precisão como $E(Y|X)$ varia com X.

Os Componentes da Variância de MQO e a Multicolinearidade

$SQT_j = \sum_{i=1}^{n} x_j^2 = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$

- Quando o tamanho da amostra cresce, cresce a variação total nos X_i. Um tamanho de amostra maior resulta em uma variância menor de $\hat{\beta}_j$.

- Esse é o componente da variância que depende do tamanho da amostra.

- LEMBRANDO: Quando SQT_j é pequeno, $Var(\hat{\beta}_j)$ pode ficar muito grande, mas um SQT_j pequeno não é violação das hipóteses do modelo..

Os Componentes da Variância de MQO e a Multicolinearidade

$(1 - R_j^2)$:

- Considere $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$.
- $Var(\hat{\beta}_1) = \frac{\sigma^2}{SQT_1(1-R_1^2)}$
- R_1^2 é o coeficiente de determinação da regressão simples de X_1 sobre X_2. Um valor de R_1^2 próximo de 1 indica que X_2 explica bastante da variação de X_1 na amostra.
- À medida que R_j^2 cresce em direção a um, $Var(\hat{\beta}_j)$ torna-se maior.

Detectando a Multicolinearidade

I. Correlação entre pares de Regressores

- Uma forma simples de detectar multicolinearidade é utilizar o coeficiente de correlação amostral, que são medidas descritivas de associação linear, entre pares de variáveis explicativas.
- Uma regra empírica: um coeficiente de correlação maior que 0.8, em valor absoluto, indica forte associação linear e uma relação de colinearidade potencialmente prejudicial.
- As relações de multicolinearidade podem envolver mais de duas variáveis explicativas, o que pode não ser detectado pelo exame de correlação dos pares.

Detectando a Multicolinearidade

II. Regressões Auxiliares

- A multicolinearidade surge devido ao fato de um ou mais regressores serem combinações lineares exatas ou aproximadas dos outros regressores.
- Uma forma de detectar a multicolinearidade e descobrir qual variável está relacionada as outras variáveis é fazer regressões auxiliares.
- A variável do membro esquerdo (dependente) é uma das variáveis explicativas e as variáveis do membro direito são todas as variáveis explicativas restantes.

$$x_{i2} = \alpha_1 x_{i1} + \alpha_3 x_{i3} + \ldots + \alpha_p x_{ip} + \nu_i$$

Detectando a Multicolinearidade

II. Regressões Auxiliares

- Caulcula-se o R_j^2 de cada uma das regressões auxiliares, que é o coeficiente de determinação na regressão da variável X_j contra as variáveis X remanescente.
- *Klein* sugere que a multicolinearidade pode representar um problema se o R_j^2 obtido de uma regressão auxiliar for superior ao R^2 da regressão de Y contra todos os regressores.
- Outro critério adotado é considerar 0.8 como ponto de corte.
- Se as associações lineares entre os regressores forem complexas, pode ser difícil identificar as inter-relações isoladas.

Detectando a Multicolinearidade

III. Fator de Inflação de Variância (FIV)

- Exprime a velocidade com o qual as variâncias e covariâncias aumentam.

$$\mathrm{FIV_j} = \frac{1}{(1 - \mathrm{R_j^2})}$$

- Assim, podemos escrever $\mathrm{Var}(\hat{\beta}_\mathrm{j}) = \frac{\sigma^2}{\sum \mathrm{x_j^2}}\mathrm{FIV_j}$

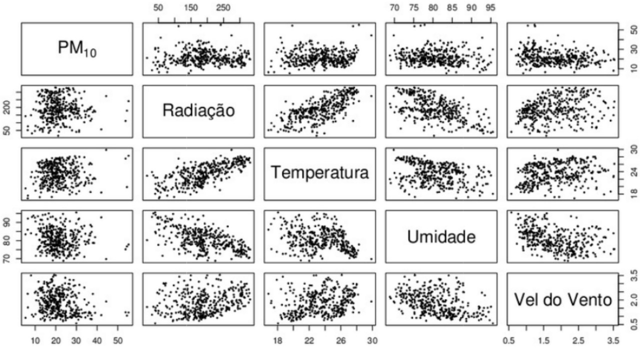
- Quando o R_j^2 aumenta, ou seja, quando a colinearidade de X_j com os outros regressores aumenta, o FIV também aumenta.

- Como regra prática, se o FIV for maior que 10 essa variável será tida como altamente colinear.

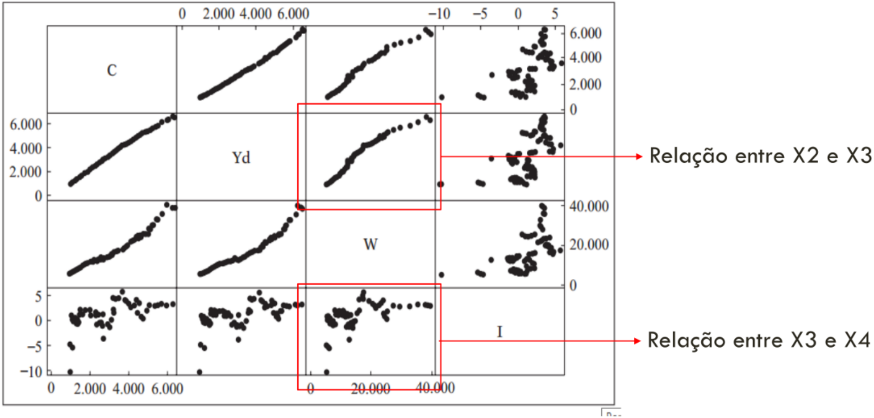
IV. Diagramas de Dispersão

Podemos observar a dispersão das variáveis.

Observações agrupadas com determinado padrão podem indicar multicolinearidade.



IV. Diagramas de Dispersão



CRÍTICA:

Não considera a influência das outras variáveis.

O que fazer se a multicolinearidade for grave?

I. O problema da multicolinearidade é que os dados não contém informação sobre os efeitos individuais das variáveis explicativas suficiente para estimar com precisão todos os parâmetros do modelo Estatístico. Uma solução consiste em aumentar a amostra.

   Essas novas informações podem constituir em dados amostrais melhores e mais numerosos, contudo, as novas observações podem sofrer da mesma relação de colinearidade e não dar grande contribuição na forma de informações novas e independentes.

O que fazer se a multicolinearidade for grave?

II. Podemos introduzir informações não amostrais sobre a forma de restrições sobre os parâmetros.

Com a utilização de restrições sobre os valores dos parâmetros, reduzimos a variabilidade amostral do estimador. Contudo, a menos que as restrições sejam exatamente verdadeiras, o estimador restrito resultante é tendencioso.

Essas informações não amostrais podem provir de princípios econômicos ou experiência anterior.

Por exemplo, poderíamos considerar que $\beta_3 = 0.2\beta_2$ ou que em um modelo para estimar demanda as variáveis explicativas: renda e preço aumentem na mesma proporção. Isso equivale a multiplicar por uma constante todos os preços e renda.

O que fazer se a multicolinearidade for grave?

III. Uma solução simples é excluir uma das variáveis colineares

Contudo, é necessário ter cautela pois podemos cometer erro de especificação devido a uma especificação incorreta do modelo.

Se a teoria econômica informa que as duas variáveis deveriam ser incluídas no modelo, excluir uma das variáveis pode constituir erro de especificação.

O que fazer se a multicolinearidade for grave?

IV. Uma solução é transformar as variáveis.

Uma possibilidade é a transformação proporcional, ou seja, o modelo transformado é obtido dividindo-se o modelo original por uma das variáveis explicativas.

Em um modelo para explicar a despesa de consumo através do PIB e população, é possível obter o modelo transformando dividindo todos os termos pela população.

Outra possibilidade se os dados são de séries temporais é fazer a regressão da primeira diferença da série.

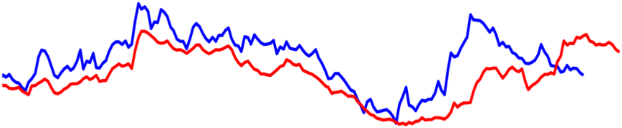
IV. Diagramas de Dispersão

- Transformação proporcional.

$$CO2_i = \hat{\beta}_1 + \hat{\beta}_2 PIB_i + \hat{\beta}_3 POP_i + \hat{u}_i$$

$$\frac{CO2}{POP_i} = \hat{\beta}_1 \left(\frac{1}{POP_i} \right) + \hat{\beta}_2 \frac{PIB_i}{POP_i} + \hat{\beta}_3 + \frac{\hat{u}_i}{POP_i}$$

- Aplicar a primeira diferença.



O que fazer se a multicolinearidade for grave?

IV. Uma solução é transformar as variáveis.

Um problema que surge da transformação proporcional é que o termo de erro do modelo transformado será heterocedástico.

E no caso da transformação de primeira diferença o termo de erro pode não satisfazer as hipóteses do modelo clássico de regressão e adcionalmente há a perda de uma observação devido ao procedimento de tomar uma diferença.

O que fazer se a multicolinearidade for grave?

V. Outro método para remediar a multicolinearidade é utilizar outras técnicas para analisar os dados.

Técnicas estatísticas multivariadas como componentes principais e análise de fator podem ser empregadas para resolver o problema.

Outra técnica bastante utilizada é a regressão ridge.

Algumas considerações

- Como a multicolinearidade não viola nenhuma das hipóteses do modelo, o problema de multicolinearidade não é de fato bem definido.

- O fato de um R_j^2 ser alto significa que x_j tem uma forte relação linear com as outras variáveis. Contudo se isso irá se traduzir em uma $\mathrm{Var}(\hat{\beta}_j)$ que é grande demais para ser útil, vai depender dos tamanhos de σ^2 e SQT_j.

- Assim, como uma valor grande de R_j^2 pode causar uma $\mathrm{Var}(\hat{\beta}_j)$ grande, uma valor pequeno de SQT_j também pode levar a variâncias grandes.

Algumas considerações

- Embora o problema de multicolinearidade não possa ser claramente definido, uma coisa é certa: tudo mais sendo igual, para estimar β_j, é melhor ter menos correlação entre x_j e as outras variáveis independentes.

- Outro ponto importante é que um elevado grau de correlação entre certas variáveis independentes pode ser irrelevante no que diz respeito a quão bem podemos estimar outros parâmetros do modelo.

- Por exemplo, se x_2 e x_3 são altamente correlacionadas mas x_1 é não correlacionada com x_2 e x_3, o valor da correlação entre x_2 e x_3 não tem efeito sobre $\mathrm{Var}(\hat{\beta}_1)$.

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SQT_1}$$

Exemplo 1: Sensibilidade dos estimadores de MQO a pequenas alterações nos dados.

Table: Dados Fictícios

Y	X_2	X_3
1	2	4
2	3	6
3	4	8
4	6	14
5	8	16

Table: Dados Fictícios Modificados

Y	X_2	X_3
1	2	4
2	3	6
3	4	14
4	6	8
5	8	16

Exemplo 1

a) Obtenha a regressão para os dados da primeira tabela e avalie a significância dos coeficientes de regressão.

b) Faça o mesmo para os dados da segunda tabela e compare a significância dos coeficientes.

c) Verifique os valores das correlações entre X_2 e X_3 nas duas regressões.

Exemplo 1

Exemplo 2

Considere o problema em economia de estimar o consumo a partir da renda e da riqueza.

Table: Dados Consumo-Renda

Y	X_2	X_3
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2052
140	220	2201
155	240	2435
150	260	2686

Exemplo 2

a) Obtenha a e equação de regressão para o problema. Avalie a significância dos parâmetros, R^2 e teste F de adequação do modelo. Todos os regressores têm sinais que atendem as expectativas?

b) Obtenha intervalos de confiança para β_2 e β_3. O que você pode concluir?

c) Faça a regressão de X_3 contra X_2. O que você pode concluir?

d) Faça as regressões de Y contra X_2 e contra X_3 separadamente. O que você pode concluir?

e) Use outras técnicas vistas em sala para detectar a presença de Multicolinearidade.

Exemplo 2

Exemplo 2

Exemplo 3

Vamos agora obter mais informações sobre os dados. Considere o banco de dados ”dadosMulticolinearidadeConsumo.R”.

a) Obtenha a e equação de regressão para o problema. Avalie a significância dos parâmetros, R^2 e teste F de adequação do modelo. Todos os regressores têm sinais que atendem as expectativas?

b) Considere agora um modelo usando a transformação logaritimica para o consumo, renda e riqueza. Repita a análise do item anterior. Nessa modelo, β_2 e β_3 dão a elasticidade da renda e riqueza e β_4 a semielasticidade.

c) Avalie os coeficientes do modelo.

d) Você acha que a multicolinearidade continua sendo um problema?

Exemplo 3

Exemplo 3